

Learning Parameters for Weighted Matrix Completion via Empirical Estimation

Jason Jo

jjo@math.utexas.edu

Mathematics Department at the University of Texas at Austin
2515 Speedway, Austin, Texas 78712

Abstract

Recently theoretical guarantees have been obtained for matrix completion in the non-uniform sampling regime. In particular, if the sampling distribution aligns with the underlying matrix's leverage scores, then with high probability nuclear norm minimization will exactly recover the low rank matrix. In this article, we analyze the scenario in which the non-uniform sampling distribution may or may not align with the underlying matrix's leverage scores. Here we explore learning the parameters for weighted nuclear norm minimization in terms of the empirical sampling distribution. We provide a sufficiency condition for these learned weights which provide an exact recovery guarantee for weighted nuclear norm minimization. It has been established that a specific choice of weights in terms of the true sampling distribution not only allows for weighted nuclear norm minimization to exactly recover the low rank matrix, but also allows for a quantifiable relaxation in the exact recovery conditions. In this article we extend this quantifiable relaxation in exact recovery conditions for a specific choice of weights defined analogously in terms of the empirical distribution as opposed to the true sampling distribution. To accomplish this we employ a concentration of measure bound and a large deviation bound. We also present numerical evidence for the healthy robustness of the weighted nuclear norm minimization algorithm to the choice of empirically learned weights. These numerical experiments show that for a variety of easily computable empirical weights, weighted nuclear norm minimization outperforms unweighted nuclear norm minimization in the non-uniform sampling regime.

1 Introduction

Matrix completion has become one of the more active fields in signal processing, enjoying numerous applications to data mining and machine learning tasks. The matrix completion problem is one where we are allowed to observe a small percentage of the entries in a data matrix \mathbf{M} and from these known entries, we must infer the values of the remaining entries. This problem is severely ill-posed, particularly so in the high dimensional regime. To this end, one must typically assume some sort of low complexity prior on \mathbf{M} , i.e. \mathbf{M} is a low rank matrix or is well approximated by a low rank matrix. Using this hypothesis a wide range of theoretical guarantees have been established for matrix completion [1, 2, 3, 6, 8, 9, 11, 12]. As noted in [4], these articles share a common thread that the recovery guarantees all require that:

- The method of sampling the data matrix \mathbf{M} must be done in a uniformly random fashion,
- And that the low-rank matrix \mathbf{M} must satisfy a so-called “incoherence” property, which roughly means that the distribution of the entries of the matrix must have some form of uniform regularity (*thereby allowing the uniform sampling strategy to be effective*).

In [4] it is observed that although the aforementioned articles differ in optimization techniques, ranging from convex relaxation via nuclear norm minimization [2], non-convex alternating minimization [8] and iterative soft thresholding [1], all of these algorithms have exact recovery guarantees using as few as $\Theta(nr \log n)$ observed elements for a square $n \times n$ matrix of rank- r .

One of the central issues in matrix completion is the relationship between the distribution of a matrix's entries and the sampling distribution being employed. For instance, if a matrix is highly incoherent, it has

much of its Frobenius norm energy spread throughout its entries in a relatively uniform fashion. To this end, taking a uniformly random sample of this matrix's entries will be a sufficient enough representation to allow for exact recovery. However, if a matrix is highly coherent, in other words, it has much of its Frobenius norm concentrated in a relatively sparse number of its entries, intuitively we understand that a uniform sampling strategy will not yield a sufficiently representative sample to allow for exact recovery.

Up until recently, the exact nature of this relationship between the \mathbf{M} and the sampling distribution \mathbf{p} has not been quantified beyond the uniform sampling case. In [4] we see this aforementioned relationship quantified. For the purposes and aims of this article, we focus on two particular results established in [4]:

- If the sampling distribution \mathbf{p} is proportional to the sum of the underlying matrix's *leverage scores*, then any arbitrary $n \times n$ rank- r matrix can be recovered from $\Theta(nr \log^2 n)$ observed entries with high probability. The exact recovery guarantee is for the nuclear norm minimization algorithm [13].
- Given a set of weights \mathbf{R}, \mathbf{C} , a sufficiency condition on the sampling distribution \mathbf{p} is established. In particular, if the sampling distribution \mathbf{p} is proportional to a sum of these \mathbf{R}, \mathbf{C} weights, then exact recovery guarantees are derived for *weighted nuclear norm minimization* (the particular form of weighted nuclear norm minimization objective was first posed in [14, 5]). Moreover, the benefit of weighted nuclear norm minimization vs. unweighted nuclear norm minimization is quantified with a specific set of weights \mathbf{R}, \mathbf{C} which are chosen in terms of the sampling distribution \mathbf{p} .

We are primarily interested in the second result on weighted nuclear norm minimization. We will explore the nature of the relationship between the weights \mathbf{R}, \mathbf{C} and the empirical sampling distribution $\hat{\mathbf{p}}$ as opposed to the true sampling distribution \mathbf{p} . As previously noted, [4] established the efficacy of weights \mathbf{R}, \mathbf{C} chosen in a specific fashion in terms of the sampling distribution \mathbf{p} . However, we are interested in a setting where the sampling distribution \mathbf{p} is not known to us and no prior knowledge of \mathbf{p} is available. In this article, we make the following contributions:

1. We extend the sufficiency condition from [4] to the case when the weights \mathbf{R}, \mathbf{C} are functions of the empirical sampling distribution $\hat{\mathbf{p}}$ for the exact recovery of \mathbf{M} using weighted nuclear norm minimization.
2. We show that a specific choice of weights \mathbf{R}, \mathbf{C} as functions of $\hat{\mathbf{p}}$ produces a similar quantifiable relaxation in exact recovery conditions for weighted nuclear norm minimization vs. unweighted nuclear norm minimization.
3. We numerically demonstrate the healthy robustness of the weighted nuclear norm minimization to the choice of the weights \mathbf{R}, \mathbf{C} , hearkening back to the previous work in non-uniform sampling and weighted matrix completion [14, 5]. We also demonstrate the superiority of weighted nuclear norm minimization over unweighted nuclear norm minimization in the non-uniform sampling regime.

To obtain the above two theoretical guarantees we will use a large deviation and a concentration of measure bound from [7] to derive sufficient conditions as to when we may use the empirical sampling distribution $\hat{\mathbf{p}}$ as an effective proxy for the true sampling distribution \mathbf{p} . The remainder of the article is organized as follow: in Section 2 we state our main results, in Section 3 we develop all the empirical estimation guarantees required to establish the matrix completion guarantees, in Section 4 we establish our matrix completion guarantees and in Section 5 we present our numerical simulations.

We use the notation that $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$ throughout the article.

2 Main Results

Numerous matrix completion results [2, 3, 12, 13] have established the effectiveness of using nuclear norm minimization:

$$\min_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{X}\|_* \text{ subject to } X_{ij} = M_{ij} \text{ for } (i, j) \in \Omega, \quad (1)$$

as a method of performing matrix completion, or in general low rank matrix recovery tasks. However, all of these results may be classified as being in the uniform sampling regime. To this end, recently [4] established

that (1) can exactly recover an $n \times n$ square matrix \mathbf{M} of rank- r from $\Theta(nr \log^2 n)$ samples as long as the sampling distribution \mathbf{p} and \mathbf{M} 's row and column leverage scores $\{\mu_i(\mathbf{M}), \nu_j(\mathbf{M})\}_{i,j=1}^n$ respectively, satisfies the following inequality:

$$p_{ij} \geq \min \left(c_0 \frac{(\mu_i(\mathbf{M}) + \nu_j(\mathbf{M}))r \log^2(2n)}{n}, 1 \right) \text{ for all } (i, j) \in [n] \times [n], \quad (2)$$

for some universal constant c_0 . With (2) the quantitative nature between the degree of non-uniformity of the sampling distribution \mathbf{p} and the corresponding coherence statistics of the matrix \mathbf{M} has been established.

Consider now a different scenario, one in which the sampling distribution \mathbf{p} and the underlying matrix's leverage scores $\{\mu_i(\mathbf{M})\}_{i=1}^{n_1}, \{\nu_j(\mathbf{M})\}_{j=1}^{n_2}$ do not align according to (2). One technique to remedy this situation is to design a transformation $\mathbf{M} \mapsto \bar{\mathbf{M}}$ so that we may adjust the leverage scores to align with the sampling distribution \mathbf{p} . Following [5, 14] we choose weights of the form $\mathbf{R} := \text{diag}(R_1, \dots, R_{n_1}) \in \mathbb{R}^{n_1 \times n_1}, \mathbf{C} := \text{diag}(C_1, \dots, C_{n_2}) \in \mathbb{R}^{n_2 \times n_2}$. Using these parameterized weights, we will use $\mathbf{M} \mapsto \mathbf{R}\mathbf{M}\mathbf{C}$ as our transformation which will adjust leverage scores of \mathbf{M} . In [14] a weighted nuclear norm objective was proposed. Following [5, 4], we will be considering the following weighted nuclear norm optimization problem:

$$\bar{\mathbf{M}} = \underset{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}}{\text{argmin}} \quad \|\mathbf{R}\mathbf{X}\mathbf{C}\|_* \text{ subject to } X_{ij} = M_{ij}, \text{ for } (i, j) \in \Omega. \quad (3)$$

In [4] exact recovery guarantees for (3) were established for weights \mathbf{R}, \mathbf{C} which were defined in terms of the true sampling distribution \mathbf{p} , which we state for the square $n \times n$ case:

Theorem 2.1. *Let $\mathbf{M} = (M_{ij})$ be an $n \times n$ matrix of rank- r , and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n] \times [n]$. Without loss of generality, assume $R_1 \leq R_2 \leq \dots \leq R_n$ and $C_1 \leq C_2 \leq \dots \leq C_n$. There exists a universal constant c_0 such that \mathbf{M} is the unique optimum to (3) with probability at least $1 - 5(2n)^{-10}$ provided that for all $(i, j) \in [n] \times [n], p_{ij} \geq n^{-10}$ and:*

$$p_{ij} \geq c_0 \left(\frac{R_i^2}{\sum_{i'=1}^{\lfloor n/(\mu_0 r) \rfloor} R_{i'}^2} + \frac{C_j^2}{\sum_{j'=1}^{\lfloor n/(\nu_0 r) \rfloor} C_{j'}^2} \right) \log^2(2n). \quad (4)$$

Note that for monotonically increasing weights \mathbf{R}, \mathbf{C} the corresponding support sets $\mathcal{S}_r, \mathcal{S}_c$ are merely the first $\lfloor n/(\mu_0 r) \rfloor$ indices, respectively.

For the remainder of the article, we shall assume that our sampling distribution \mathbf{p} has a product form $p_{ij} = p_i^r p_j^c$ for all $(i, j) \in [n_1] \times [n_2]$. Furthermore, we will consider the following *two-stage sampling model*:

- Stage 1 (Empirical Sampling Distribution): We sample the distribution \mathbf{p} with m times independently with replacement, but the corresponding entries of the data matrix \mathbf{M} are not revealed to us. In other words, we are *sampling the sampling distribution*, but not the underlying matrix \mathbf{M} .
- Stage 2 (Sampling the Matrix): We then, independent of the first stage, sample the matrix \mathbf{M} using the independent Bernoulli model for each entry $(i, j) \in [n_1] \times [n_2]$.

Note that this two stage sampling models allows one to sample the sampling distribution \mathbf{p} without revealing the entries of \mathbf{M} . In this manner we may design weights \mathbf{R}, \mathbf{C} which depend on the empirical sampling distribution $\hat{\mathbf{p}}$ and obtain matrix completion guarantees for these weights in the usual (stage two) independent Bernoulli sampling model that has been typically used in the matrix completion literature.

In this article we present stage one sampling bounds which will allow $\hat{\mathbf{p}}$ to be used as an empirical proxy for \mathbf{p} to design weights \mathbf{R}, \mathbf{C} for (3) and obtain exact recovery with high probability. To this end, we establish the following two empirical estimation lemmas, which will serve as the foundation to our matrix completion guarantees. The first is a *one sided large deviation bound*:

Lemma 2.2. *Let \mathbf{p} denote a probability mass function on $[n_1] \times [n_2]$ and suppose \mathbf{p} has a product form, i.e. for all $(i, j) \in [n_1] \times [n_2] : p_{ij} = p_i^r p_j^c$ for $\mathbf{p}^r, \mathbf{p}^c$ probability mass functions on $[n_1], [n_2]$, respectively. Let $X_1, \dots, X_m \stackrel{i.i.d}{\sim} \mathbf{p}$ be a sequence of m i.i.d samples. For any $\alpha \in (0, (\min_{i \in [n_1]} p_i^r \vee \min_{j \in [n_2]} p_j^c)^{-1})$ and $\epsilon \in (0, 1)$, if the number of samples m is chosen such that:*

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c \right)^{-2} \log(\epsilon^{-1}(n_1 + n_2)), \quad (5)$$

then with probability at least $1 - \epsilon$ we have that for all $(i, j) \in [n_1] \times [n_2]$:

$$p_{ij} \geq \frac{1}{(1 + \alpha)^2} \hat{p}_{ij}. \quad (6)$$

We also establish the following *two sided empirical bound* for the estimation of product distributions:

Lemma 2.3. *Let \mathbf{p} denote a probability mass function on $[n_1] \times [n_2]$ and suppose \mathbf{p} has a product form, i.e. for all $(i, j) \in [n_1] \times [n_2]$: $p_{ij} = p_i^r p_j^c$ for $\mathbf{p}^r, \mathbf{p}^c$ probability mass functions on $[n_1], [n_2]$, respectively. Let $X_1, \dots, X_m \stackrel{i.i.d}{\sim} \mathbf{p}$ be a sequence of m i.i.d samples. For any $\alpha \in (0, (\min_{i \in [n_1]} p_i^r \vee \min_{j \in [n_2]} p_j^c)^{-1})$ and $\epsilon \in (0, 1)$, if the number of samples m is chosen such that:*

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c \right)^{-2} \log(2\epsilon^{-1}(n_1 + n_2)), \quad (7)$$

then with probability at least $1 - \epsilon$ we have that for all $(i, j) \in [n_1] \times [n_2]$:

$$\frac{1}{(1 + \alpha)^2} \hat{p}_{ij} \leq p_{ij} \leq \frac{1}{(1 - \alpha)^2} \hat{p}_{ij}. \quad (8)$$

Note that Lemmas 2.2 and 2.3 are general results for the empirical estimation of any distribution \mathbf{p} over $[n_1] \times [n_2]$ which has a product form. Recall that the sampling model employed in [4] is a sequence of $n_1 \cdot n_2$ independent Bernoulli random variables, with each Bernoulli random variable having success probability p_{ij} for $(i, j) \in [n_1] \times [n_2]$. Therefore, \mathbf{p} may not be a probability matrix on $[n_1] \times [n_2]$ as it may not sum to 1. To this end, we note that when we sample \mathbf{p} , we are really sampling the normalized matrix $\frac{1}{\sum_{i,j} p_{ij}} \mathbf{p}$. So our empirical estimator $\hat{\mathbf{p}}$ is estimating the normalized probability matrix $\frac{1}{\sum_{i,j} p_{ij}} \mathbf{p}$ and not \mathbf{p} itself. Therefore, in order to apply the above lemmas we must account for this normalization constant.

Using the above, we will obtain two weighted matrix completion guarantees. For simplicity, we will prove all our results for the case when \mathbf{M} is a square $n \times n$ matrix. The first guarantee will be a sufficiency condition for the weights \mathbf{R}, \mathbf{C} in terms of the empirical estimator $\hat{\mathbf{p}}$ which will ensure exact recovery by weighted nuclear norm minimization with high probability:

Theorem 2.4. *Let $\mathbf{M} = (M_{ij})$ be an $n \times n$ matrix of rank- r , and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n] \times [n]$, Let $\epsilon \in (0, 1)$ be arbitrary. Suppose that there exists an $\alpha \in (0, (\min_{i \in [n]} p_i^r / (\sum_{i \in [n]} p_i^r) \vee \min_{j \in [n]} p_j^c / (\sum_{j \in [n]} p_j^c))^{-1})$ and some universal constant c_0 such that for all indices $(i, j) \in [n] \times [n]$ the weights \mathbf{R}, \mathbf{C} satisfy the following inequalities:*

$$\hat{p}_{ij} \geq \frac{(1 + \alpha)^2}{\sum_{ij} p_{ij}} c_0 \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n), \quad (9)$$

where $\mathcal{S}_r, \mathcal{S}_c$ denote the $\lfloor n/(\mu_0 r) \rfloor$ entries of least magnitude of \mathbf{R}, \mathbf{C} , respectively. If the number of stage one samples m is chosen such that:

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j \in [n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c} \right)^{-2} \log(2\epsilon^{-1}n)$$

and if for all $(i, j) \in [n] \times [n], p_{ij} \geq n^{-10}$, then with probability at least $(1 - 5(2n)^{-10})(1 - \epsilon)$, \mathbf{M} is unique optimum to (3), where Ω is obtained via the usual (stage two) independent, entry-wise Bernoulli sampling of \mathbf{M} .

Our second weighted matrix completion guarantee will be for the exact recovery properties of a set weights \mathbf{R}, \mathbf{C} explicitly defined in terms of the empirical distribution $\hat{\mathbf{p}}$:

Theorem 2.5. *Let \mathbf{M} be a square $n \times n$ rank- r matrix with coherence μ_0 . Consider the weights defined by:*

$$R_i = \sqrt{\frac{1}{n} \hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c}, \text{ for } i = 1, \dots, n, \quad (10)$$

$$C_j = \sqrt{\frac{1}{n} \hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r} \text{ for } j = 1, \dots, n, \quad (11)$$

where $\mathcal{S}_r, \mathcal{S}_c$ denote the $\lfloor n/(\mu_0 r) \rfloor$ entries of $\hat{\mathbf{p}}^r, \hat{\mathbf{p}}^c$ of least magnitude, respectively. Suppose that there exists an $\alpha \in (0, (\min_{i \in [n]} p_i^r / (\sum_{i \in [n]} p_i^r) \vee \min_{j \in [n]} p_j^c / (\sum_{j \in [n]} p_j^c))^{-1})$ such that the (unnormalized) matrix \mathbf{p} satisfies for all $(i, j) \in [n] \times [n]$ and the sets $\mathcal{S}_r^*, \mathcal{S}_c^*$ which denote the $\lfloor n/(\mu_0 r) \rfloor$ entries of $\mathbf{p}^r, \mathbf{p}^c$ of least magnitude, respectively satisfies the following:

$$p_j^c \sum_{i' \in \mathcal{S}_r^*} p_{i'}^r \geq c_0 \frac{2(1+\alpha)^2}{(1-\alpha)^2} \log^2(2n), \quad (12)$$

$$p_i^r \sum_{j' \in \mathcal{S}_c^*} p_{j'}^c \geq c_0 \frac{2(1+\alpha)^2}{(1-\alpha)^2} \log^2(2n). \quad (13)$$

If the number of stage one samples m is chosen such that:

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j \in [n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c} \right)^{-2} \log(4\epsilon^{-1}n),$$

then with probability at least $(1 - 5(2n)^{-10})(1 - \epsilon)$, \mathbf{M} is unique optimum to (3), where Ω is obtained via the usual (stage two) independent, entry-wise Bernoulli sampling of \mathbf{M} .

Note: Unweighted nuclear norm minimization attains exact recovery under the condition that for all $(i, j) \in [n] \times [n]$:

$$p_i^r p_j^c \gtrsim \frac{\mu_0 r}{n} \log^2(2n). \quad (14)$$

However as Theorem 2.5 establishes, weighted nuclear norm minimization with choice of weights (10) and (11) attains exact recovery subject to the less restrictive sufficient recovery condition that:

$$p_j^c \sum_{i' \in \mathcal{S}_r^*} p_{i'}^r \gtrsim \log^2(2n),$$

$$p_i^r \sum_{j' \in \mathcal{S}_c^*} p_{j'}^c \gtrsim \log^2(2n).$$

This is precisely the condition from [4].

3 Empirical Estimation

We consider probability mass functions \mathbf{p} on $[n_1] \times [n_2]$ which have a product form $p_{ij} = p_i^r p_j^c$ for $(i, j) \in [n_1] \times [n_2]$. We will sample this distribution with replacement m times. The $X_1, \dots, X_m \stackrel{i.i.d}{\sim} \mathbf{p}$ samples are row and column pairs, i.e. $X_k \in [n_1] \times [n_2]$ for each $k = 1, \dots, m$. We may define the *row and column empirical estimators*:

Definition 3.1. The row and column empirical estimators $\hat{\mathbf{p}}^r, \hat{\mathbf{p}}^c$, respectively are defined as:

$$\hat{p}_i^r := \frac{1}{m} \sum_{k=1}^m \delta_i^r(X_k), \text{ for } i \in [n_1], \quad (15)$$

$$\hat{p}_j^c := \frac{1}{m} \sum_{k=1}^m \delta_j^c(X_k), \text{ for } j \in [n_2], \quad (16)$$

where for any X_k :

$$\delta_i^r(X_k) = \begin{cases} 1 & \text{if } X_k \text{ is from row } i, \\ 0 & \text{otherwise.} \end{cases}$$

$$\delta_j^c(X_k) = \begin{cases} 1 & \text{if } X_k \text{ is from column } j, \\ 0 & \text{otherwise.} \end{cases}$$

For the remainder of the article, we will allow $\hat{\mathbf{p}}$ denote the empirical product estimate, i.e. $\hat{\mathbf{p}} = \hat{\mathbf{p}}^r \hat{\mathbf{p}}^c$.

Observe that in (15) and (16) each component of our row and column empirical estimators involve a sum of independent, bounded in $[0, 1]$ random variables as $\delta_i^r(X_k), \delta_j^c(X_k) \in \{0, 1\}$ for any $(i, j, k) \in [n_1] \times [n_2] \times [m]$. In this situation, we may use *Hoeffding's inequalities* [7] to obtain some probabilistic approximation guarantees. For our purposes, we will be using two forms of Hoeffding's inequalities: a one sided large deviation bound and a two sided concentration of measure bound.

Theorem 3.2. (*Hoeffding Inequalities*) Let Z_1, \dots, Z_m be independent random variables such that each $Z_i \in [a_i, b_i]$ with probability 1. Let $S_m = \sum_{i=1}^m Z_i$. Then for any $t > 0$ we have:

$$\Pr[S_m - \mathbb{E}[S_m] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right), \quad (17)$$

$$\Pr[|S_m - \mathbb{E}[S_m]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right). \quad (18)$$

For any $i \in [n_1]$, we may define m random variables $Z_{i,k} := \delta_i^r(X_k)$ for $k = 1, \dots, m$. Note that each random variable $Z_{i,k}$ only takes values in $\{0, 1\}$ and thus is bounded in $[0, 1]$ with probability 1. As each X_k is merely a row and column index, and each δ_i^r, δ_j^c are row and column indicator functions, we have that any set of the $Z_{i,k}$'s (and similarly for the column case) is an independent set of random variables. Therefore the hypotheses of Theorem 3.2 are satisfied. For each $i \in [n_1]$ we may define the sum $S_{i,m}^r := \sum_{k=1}^m Z_{i,k}$. Each $S_{i,m}^r$ has expected value $\mathbb{E}[S_{i,m}^r] = mp_i^r$. Analogous results hold for the column case. With the above pair of Hoeffding inequalities in hand, we are now ready to establish our main lemmas. For the proof of Lemma 2.2 we will apply (17) and for the proof of Lemma 2.3 we will apply (18).

3.1 Proof Lemma 2.2

Proof. We start our proof by analyzing empirical estimation of the row distribution; the analysis for the column distribution will be identical. For any $i \in [n_1], \alpha > 0$, choosing $t = \alpha \min_{i \in [n_1]} p_i^r$, by (17) we have that:

$$\Pr[\hat{p}_i^r - p_i^r \geq \alpha \min_{i \in [n_1]} p_i^r] \leq \exp(-2(\alpha \min_{i \in [n_1]} p_i^r)^2 m). \quad (19)$$

We may repeat the analysis for the column case, where we choose $t = \alpha \min_{j \in [n_2]} p_j^c$, then analogously:

$$\Pr[\hat{p}_j^c - p_j^c \geq \alpha \min_{j \in [n_2]} p_j^c] \leq \exp(-2(\alpha \min_{j \in [n_2]} p_j^c)^2 m). \quad (20)$$

For any $i \in [n_1]$ let E_i^r denote the event that $\hat{p}_i^r - p_i^r \geq \alpha \min_{i \in [n_1]} p_i^r$ and for any $j \in [n_2]$ let E_j^c denote the event that $\hat{p}_j^c - p_j^c \geq \alpha \min_{j \in [n_2]} p_j^c$.

We must choose $\alpha > 0$ such that the bounds in (19), (20) are nontrivial. In particular, any two probability vectors cannot have their components differ by more than 1. Therefore, we require that α satisfies:

$$\alpha \min_{i \in [n_1]} p_i^r \leq 1 \text{ and } \alpha \min_{j \in [n_2]} p_j^c \leq 1.$$

To this end it suffices to choose $\alpha \in (0, (\min_{i \in [n_1]} p_i^r \vee \min_{j \in [n_2]} p_j^c)^{-1})$.

By (19), (20) and the Union Bound we have that:

$$\begin{aligned} \Pr[\text{For some } (i, j) \text{ the event } E_i^r \text{ or } E_j^c \text{ occurs}] &\leq \left(n_1 \exp(-2(\alpha \min_{i \in [n_1]} p_i^r)^2 m) + n_2 \exp(-2(\alpha \min_{j \in [n_2]} p_j^c)^2 m) \right) \\ &\leq (n_1 + n_2) \exp(-2(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c)^2 m). \end{aligned} \quad (21)$$

Observe that (21) immediately yields that with probability at least $1 - (n_1 + n_2) \exp(-2(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c)^2 m)$ for any $(i, j) \in [n_1] \times [n_2]$ we have that the following bounds hold:

$$\hat{p}_i^r - p_i^r \leq \alpha \min_{i \in [n_1]} p_i^r, \quad (22)$$

$$\hat{p}_j^c - p_j^c \leq \alpha \min_{j \in [n_2]} p_j^c. \quad (23)$$

Therefore with probability at least $1 - (n_1 + n_2) \exp(-2(\alpha \min_{i,j} p_{i,j})^2 m)$ we may conclude that for all $(i, j) \in [n_1] \times [n_2]$ the following bound is true:

$$p_{ij} \geq \frac{1}{(1 + \alpha)^2} \hat{p}_{ij}. \quad (24)$$

For any $\epsilon \in (0, 1)$ choosing m such that:

$$m = \frac{\log((n_1 + n_2)\epsilon^{-1})}{2 \left(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c \right)^2}, \quad (25)$$

guarantees that (24) holds with probability at least $(1 - \epsilon)$ and the proof is complete. \square

3.2 Proof of Lemma 2.3

Proof. The proof of Lemma 2.3 is similar to the previous proof but we include the full proof for completeness. We start our proof by analyzing empirical estimation of the row distribution; the analysis for the column distribution will be identical. Following the previous section we restrict ourselves to choose $\alpha \in (0, (\min_{i \in [n_1]} p_i^r \vee \min_{j \in [n_2]} p_j^c)^{-1})$. For any $i \in [n_1]$ choosing $t = \alpha \min_{i \in [n_1]} p_i^r$, by (18) we have that:

$$\Pr[|\hat{p}_i^r - p_i^r| \geq \alpha \min_{i \in [n_1]} p_i^r] \leq 2 \exp(-2(\alpha \min_{i \in [n_1]} p_i^r)^2 m). \quad (26)$$

We may repeat the analysis for the column case, where we choose $t = \alpha \min_{j \in [n_2]} p_j^c$, then analogously:

$$\Pr[|\hat{p}_j^c - p_j^c| \geq \alpha \min_{j \in [n_2]} p_j^c] \leq 2 \exp(-2(\alpha \min_{j \in [n_2]} p_j^c)^2 m). \quad (27)$$

For any $i \in [n_1]$ let E_i^r denote the event that $|\hat{p}_i^r - p_i^r| \geq \alpha \min_{i \in [n_1]} p_i^r$ and for any $j \in [n_2]$ let E_j^c denote the event that $|\hat{p}_j^c - p_j^c| \geq \alpha \min_{j \in [n_2]} p_j^c$. By (26), (27) and the Union Bound we have that:

$$\begin{aligned} \Pr[\text{For some } (i, j) \text{ the event } E_i^r \text{ or } E_j^c \text{ occurs}] &\leq 2 \left(n_1 \exp(-2(\alpha \min_{i \in [n_1]} p_i^r)^2 m) + n_2 \exp(-2(\alpha \min_{j \in [n_2]} p_j^c)^2 m) \right) \\ &\leq 2(n_1 + n_2) \exp(-2(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c)^2 m). \end{aligned} \quad (28)$$

Observe that (28) immediately yields that with probability at least $1 - 2(n_1 + n_2) \exp(-2(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c)^2 m)$ for any $(i, j) \in [n_1] \times [n_2]$ we have that the two following bounds hold:

$$|\hat{p}_i^r - p_i^r| \leq \alpha \min_{i \in [n_1]} p_i^r, \quad (29)$$

$$|\hat{p}_j^c - p_j^c| \leq \alpha \min_{j \in [n_2]} p_j^c. \quad (30)$$

The bound (29) is equivalent to the following:

$$-\alpha \min_{i \in [n_1]} p_i^r \leq \hat{p}_i^r - p_i^r \leq \alpha \min_{i \in [n_1]} p_i^r,$$

and the above inequality yields that for any $i \in [n_1]$:

$$\frac{1}{1 + \alpha} \hat{p}_i^r \leq p_i^r \leq \frac{1}{1 - \alpha} \hat{p}_i^r. \quad (31)$$

Similarly (30) implies that for any $j \in [n_2]$:

$$\frac{1}{1 + \alpha} \hat{p}_j^c \leq p_j^c \leq \frac{1}{1 - \alpha} \hat{p}_j^c. \quad (32)$$

Combining (31) and (32), we have that with probability at least $1 - 2(n_1 + n_2) \exp(-2(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c)^2 m)$ that:

$$\frac{1}{(1 + \alpha)^2} \hat{\mathbf{p}} \leq \mathbf{p} \leq \frac{1}{(1 - \alpha)^2} \hat{\mathbf{p}}. \quad (33)$$

For any $\epsilon \in (0, 1)$ note that if we choose:

$$m = \frac{\log(2(n_1 + n_2)\epsilon^{-1})}{2(\alpha \min_{i \in [n_1]} p_i^r \wedge \min_{j \in [n_2]} p_j^c)^2}, \quad (34)$$

then (33) holds with probability at least $1 - \epsilon$ and the proof is complete. \square

4 Matrix Completion Guarantees

With Lemma 2.2 in hand, we are prepared to prove Theorem 2.4 in Section 4.1. In Section 4.2, using Lemma 2.3 we will prove Theorem 2.5 which quantifies the relaxation of the condition for which (3) succeeds in obtaining exact recovery using the empirically learned weights when compared to unweighted nuclear norm minimization.

4.1 Proof of Theorem 2.4

Proof. For any $\alpha \in (0, (\min_{i \in [n]} p_i^r / (\sum_{i=1}^n p_i^r) \vee \min_{j \in [n]} p_j^c / (\sum_{j=1}^n p_j^c))^{-1})$ and $\epsilon \in (0, 1)$ if we choose

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j \in [n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c} \right)^{-2} \log(2\epsilon^{-1}n)$$

by Lemma 2.2 we have that with probability at least $(1 - \epsilon)$ for any $(i, j) \in [n] \times [n]$:

$$\frac{p_{ij}}{\sum_{ij} p_{ij}} \geq \frac{1}{(1 + \alpha)^2} \hat{p}_{ij}. \quad (35)$$

Observe that if the weights \mathbf{R}, \mathbf{C} satisfy (9) for α , we have that:

$$p_{ij} \geq \frac{\sum_{ij} p_{ij}}{(1 + \alpha)^2} \hat{p}_{ij} \quad (36)$$

$$\geq c_0 \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n). \quad (37)$$

By Theorem 2.1 (37) is sufficient to guarantee exact recovery of \mathbf{M} via (3) with probability at least $1 - 5(2n)^{-10}$. As stage one and stage two sampling are independent, we conclude that (3) attains exact recovery with probability at least $(1 - 5(2n)^{-10})(1 - \epsilon)$. \square

4.2 Weighted Nuclear Norm and Relaxation of Sufficient Recovery Conditions

With Theorem 2.4 we established some sufficient conditions for the weights \mathbf{R}, \mathbf{C} in order for (3) to attain exact recovery. In this section we will establish exact recovery guarantees for a specific set of weights defined in terms of the empirical sampling distribution $\hat{\mathbf{p}}$ and quantify how the exact recovery conditions for (3) are relaxed relative to unweighted nuclear norm minimization (1).

4.2.1 Proof of Theorem 2.5

Proof. Choosing the weights \mathbf{R}, \mathbf{C} as in (10) and (11), observe that for any $(i, j) \in [n] \times [n]$:

$$\left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n) = \left(\frac{\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c + \hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r}{\sum_{i', j' \in \mathcal{S}_r, \mathcal{S}_c} \hat{p}_{i'}^r \hat{p}_{j'}^c} \right) \log^2(2n). \quad (38)$$

Let $\alpha \in (0, (\min_{i \in [n]} p_i^r \vee \min_{j \in [n_2]} p_j^c)^{-1})$ be such that (12) and (13) hold and let $\epsilon \in (0, 1)$ be arbitrary. By Lemma 2.3 choosing m such that:

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j \in [n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c} \right)^{-2} \log(4\epsilon^{-1}n)$$

guarantees that with probability at least $(1 - \epsilon)$ that for all indices $(i, j) \in [n] \times [n]$:

$$\frac{1}{(1 + \alpha)^2} \hat{p}_{ij} \leq \frac{p_{ij}}{\sum_{i,j} p_{ij}} \leq \frac{1}{(1 - \alpha)^2} \hat{p}_{ij}. \quad (39)$$

Applying (39) to (38) we have that for any $(i, j) \in [n] \times [n]$:

$$\begin{aligned} \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n) &= \left(\frac{\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c + \hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r}{\sum_{i', j' \in \mathcal{S}_r, \mathcal{S}_c} \hat{p}_{i'}^r \hat{p}_{j'}^c} \right) \log^2(2n) \\ &\leq \frac{(1 + \alpha)^2}{(1 - \alpha)^2} \left(\frac{p_i^r \sum_{j' \in \mathcal{S}_c} p_{j'}^c + p_j^c \sum_{i' \in \mathcal{S}_r} p_{i'}^r}{\sum_{i', j' \in \mathcal{S}_r, \mathcal{S}_c} p_{i'}^r p_{j'}^c} \right) \log^2(2n) \\ &= \frac{(1 + \alpha)^2}{(1 - \alpha)^2} \left[\frac{p_i^r \log^2(2n)}{\sum_{i' \in \mathcal{S}_r} p_{i'}^r} + \frac{p_j^c \log^2(2n)}{\sum_{j' \in \mathcal{S}_c} p_{j'}^c} \right] \\ &\leq \frac{(1 + \alpha)^2}{(1 - \alpha)^2} \left[\frac{p_i^r \log^2(2n)}{\sum_{i' \in \mathcal{S}_r^*} p_{i'}^r} + \frac{p_j^c \log^2(2n)}{\sum_{j' \in \mathcal{S}_c^*} p_{j'}^c} \right] \end{aligned} \quad (40)$$

$$\leq \frac{1}{c_0} p_{ij}. \quad (41)$$

where (40) follows as the sets $\mathcal{S}_r^*, \mathcal{S}_c^*$ serve as a lower bound for the terms $\sum_{i' \in \mathcal{S}_r} p_{i'}^r, \sum_{j' \in \mathcal{S}_c} p_{j'}^c$, respectively and thus inverting they serve as an upper bound and (41) follows from (12) and (13). Again by Theorem 2.1 we immediately see that (41) is sufficient to guarantee exact recovery of \mathbf{M} via (3) with probability at least $1 - 5(2n)^{-10}$. \square

5 Numerical Experiments

Here we test the performance of weighted nuclear norm minimization using various weights. We have the following experimental setup: the data matrix \mathbf{M} is a unit Frobenius norm standard normal Gaussian square matrix of dimension $n = 500$. Our sampling distribution $\mathbf{p} = \mathbf{p}^r \mathbf{p}^c$ where $\mathbf{p}^r, \mathbf{p}^c$ are power law distributed with exponent equal to 1.2. Sampling the distribution \mathbf{p} at a rate of m times with replacement and we obtain the empirical distribution $\hat{\mathbf{p}} = \hat{\mathbf{p}}^r \hat{\mathbf{p}}^c$. Using this empirical distribution $\hat{\mathbf{p}}$ we test nuclear norm minimization using the following weights, as was done in [5]:

1. Unweighted (Uniform Weights): the weights \mathbf{R}, \mathbf{C} are equal to the uniform weights.
2. True Weighted: the weights \mathbf{R}, \mathbf{C} satisfy: $\mathbf{R} = (\mathbf{p}^r)^{1/2}, \mathbf{C} = (\mathbf{p}^c)^{1/2}$.
3. Empirically Weighted: the weights \mathbf{R}, \mathbf{C} satisfy: $\mathbf{R} = (\hat{\mathbf{p}}^r)^{1/2}, \mathbf{C} = (\hat{\mathbf{p}}^c)^{1/2}$.
4. Empirically Smoothed Weights: the weights \mathbf{R}, \mathbf{C} are a linear combination of the empirical weights and the uniform weights. Letting $\mathbf{1}_n := [1, \dots, 1]$ be a vector of length n whose coordinates are all equal to 1, we set $\mathbf{R} = \frac{1}{2n} (\hat{\mathbf{p}}^r)^{1/2} + \frac{1}{2n} \mathbf{1}_n$ and $\mathbf{C} = \frac{1}{2} (\hat{\mathbf{p}}^c)^{1/2} + \frac{1}{2n} \mathbf{1}_n$, i.e. we put half of the mass on the empirical distribution and remaining half of the mass on the uniform weights.

We let the rank of \mathbf{M} be 5, 10, 15, 20, 25 and we choose a range of variable sampling rates. For each rank and sampling rate test configuration we performed 100 trials. We consider exact recovery to be when the output of the weighted nuclear norm $\bar{\mathbf{M}}$ satisfies: $\|\mathbf{M} - \bar{\mathbf{M}}\|_F \leq 10^{-5}$. To execute the weighted nuclear

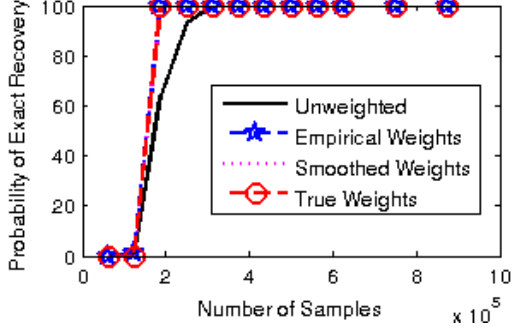


Figure 1: Probability of Exact Recovery when the rank is equal to 5.

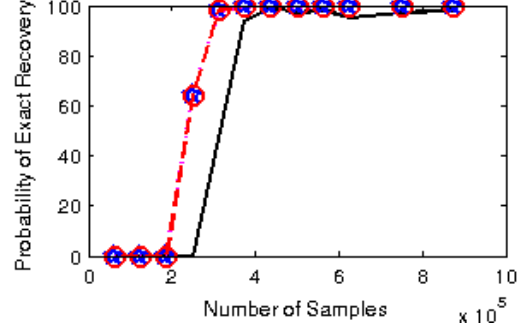


Figure 2: Probability of Exact Recovery when the rank is equal to 10.

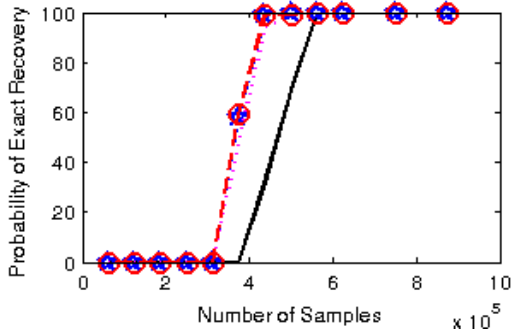


Figure 3: Probability of Exact Recovery when the rank is equal to 15.

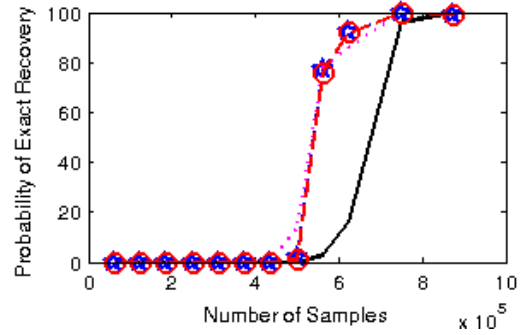


Figure 4: Probability of Exact Recovery when the rank is equal to 20.

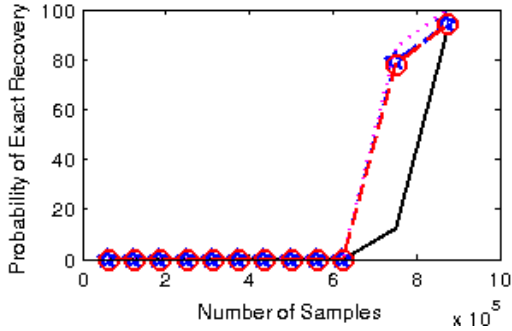


Figure 5: Probability of Exact Recovery when the rank is equal to 25.

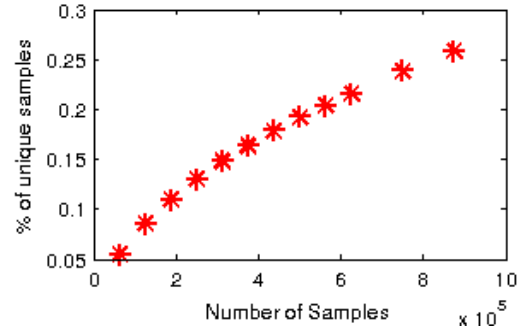


Figure 6: Power Law Sampling with replacement rate vs. Percentage of Unique Samples Revealed.

norm minimization program we utilized the Augmented Lagrangian Method [10]. We obtained the following phase transition diagrams in Figures 1-5.

Note that we do not perform the two stage sampling method. As the power law sampling distribution \mathbf{p} is non-uniform, even though we may sample at a rate of $m = O(n_1 n_2)$, the rate that the percentage of unique revealed entries of \mathbf{M} grows is in line with the uniform sampling regime we are accustomed to. In Figure 6 we show how with the independent sampling with replacement rate m grows with the percentage of unique entries of \mathbf{M} .

6 Conclusion

In this article we extended numerous weighted nuclear norm minimization results from [4]. In particular we extended results where the weights were being defined in relation to the true sampling distribution \mathbf{p} to the weights being defined in relation to the empirical sampling distribution $\hat{\mathbf{p}}$. Furthermore, we defined an empirical set of weights and established a quantifiable relaxation of exact recovery conditions for weighted nuclear norm minimization when compared to the unweighted nuclear norm. To achieve these guarantees we used a large deviation bound and a concentration of measure inequality from [7]. We showed that weighted nuclear norm minimization is quite robust to the choice of empirically learned weights. Indeed, we used a broad range of empirical weights and saw strikingly similar exact recovery gains over unweighted nuclear norm minimization.

7 Acknowledgements

The author would like to acknowledge and thank Rachel Ward for their insight and guidance throughout this project.

References

- [1] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4):1956–1982, March 2010.
- [2] Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, June 2012.
- [3] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, May 2010.
- [4] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing Any Low-rank Matrix, Provably. *ArXiv e-prints arXiv:1306.2979v4*, 2014.
- [5] Rina Foygel, Ruslan Salakhutdinov, Ohad Shamir, and Nati Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. *NIPS Proceedings*, 24, 2011.
- [6] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theor.*, 57(3):1548–1566, March 2011.
- [7] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [8] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC ’13, pages 665–674, 2013.
- [9] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.
- [10] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.
- [11] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13.
- [12] Benjamin Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, December 2011.
- [13] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, August 2010.

- [14] Ruslan Salakhutdinov and Nathan Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. arxiv.org/abs/1002.2780, 2010.